04 | 数据分析：数据的价值有哪些？

2020-11-16 郭朝斌

物联网开发实战 进入课程 >

讲述：郭朝斌
时长 18:26 大小 16.88M

你好，我是郭朝斌。

在前两讲，我们了解了物联网设备层的通信技术和网络层的网络协议。设备借助通信技术，能接入网络；而统一的网络协议，保证了云平台能够“看懂”这些上报的数据。

但是，如果只是单纯地采集和存储的话，那么数据就只是数据库中的 0 和 1 ，很难谈得上有什么价值。某种程度上说，物联网系统的价值其实就在于**数据的价值**，而数据的价值则来源于我们对数据的分析和应用。

所以这一讲，我们来到了物联网数据流的第三站，应用层。开始下面的学习之前，你可以假设一下，如果你是物联网平台开发人员，你会怎么有效地分析和应用数据呢？

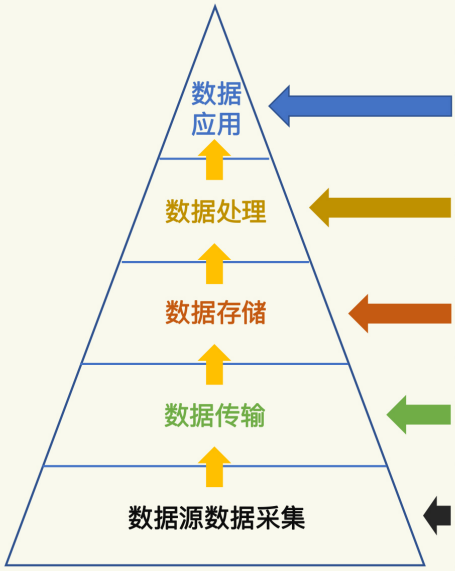
从数据的角度审视物联网

回答这个问题之前，我想先从数据的角度，带你对物联网体系结构再做一次梳理。

你可能有疑惑了，之前我们整理的物联网三大层面，不是也跟数据流向有关吗？是的，但是那三个层面是基于各部分的特性划分的，包含的技术内容更广。

而今天这一讲，我只专注数据这一个维度，对数据从“生产”到“消费”的过程做一个更细化的拆分和整理，得出一个技术体系。通过这个体系，你就能够对数据相关的技术有一个更加清晰的认识，而且学习过程中心里时刻有个宏观视角，不会“迷失”在各种技术的琐碎细节里面。

我先给出这张物联网数据分析的技术体系示意图，然后我会结合这张图片，从下往上分别讲解其中的每一步：



相关技术	关键点
数据可视化、数据挖掘、商业智能、数据预测、控制决策	价值和算法
批处理、流处理	高效和简单
结构化数据：分布式数据库 半结构数据：NoSQL等 非结构数据：HDFS等	海量数据
消息队列、负载均衡等	高并发
嵌入式系统、传感器	多样性

第一步：数据源数据采集

数据源数据采集，就是通过各种物联网设备，实现各种数据数字化的任务，同时也可能需要将数据临时存储起来，准备好数据的上传工作。

因为数据类型多种多样，比如我们常说的温度、湿度、甲醛含量，另外还有图像、声音、震动等等，所以采集这些数据的设备也具有明显的**多样性**。

除了传统的**嵌入式系统**涉及的软件和硬件开发外，**传感器**是这一部分需要重点关注的关键元器件。随着物联网的发展，新的应用在不断涌现，比如监测人体心率和血氧饱和度的可穿戴设备，以及在智慧农业中用来测量土壤反射率和植物颜色的光学传感器等等。

传感器是一个具有技术门槛的行业，高度依赖**材料技术**的进步。除了半导体和光纤这些相对传统的材料，现在一些纳米和生物材料的应用也发展很快。

采集完数据后，你就要利用网络将数据传输到云服务器。

第二步：数据传输

数据传输，就是把采集的数据快速地、可靠地上报到云平台。

这里的技术难点是实现**高并发**。因为物联网设备的数量巨大，而且它们在时刻不断地产生数据，所以海量的数据对云平台的性能构成了挑战。云平台只有具备高并发的能力，才能有效地、可靠地实现数据的传输。

高并发的实现，需要借助**分布式**的架构设计，同时使用**负载均衡**、**消息队列**和**缓存**等相关技术。我将在进阶篇针对这个话题再展开讲一讲。

第三步：数据存储

数据传输完成之后，你要面对问题是**海量数据的存储**。

比如，一辆共享单车一天会产生 1MB 的数据，如果按照 2000 万辆的保有量计算，每天就会产生 20TB 的新增数据。而在智慧城市的系统中，一个 800 万像素的摄像头，一个小时就能产生 3.6GB 的数据。那么一座城市，光是摄像头，一个月的数据量就可以达到数百 PB。

除了数据量大，数据的种类也很丰富，它们大体上可以分为 3 类，分别需要有不同的存储手段：

1. **结构化数据**，比如用户和设备的关系，用户信息、设备参数等。这类数据还是适合**关系型数据库**，那为了应对海量数据，你可以采用**分布式数据库**。另外，物联网中的传感器

设备，随时间不断产生新数据。要存储这类数据，你可以选择**时序数据库**，来获得更高的读写和查询性能。

2. **半结构化数据**，比如 JSON 结构的数据，日志记录等。这些数据的存储一般采用**NoSQL 数据库**产品。
3. **非结构化数据**，比如视频、音频等数据。它们一般采用文件的形式存储，Hadoop 体系中应用广泛的**分布式文件系统 HDFS** 是常用的选项。

关于数据存储的详细内容，我会在进阶篇为你进行剖析。

只有当实现了海量数据的存储，我们也才能为数据的处理打好基础。

第四步：数据处理

数据处理阶段，非常考验你对海量数据的处理能力，你需要用到**大数据计算引擎**。大数据的处理，从应用场景划分的话，可以分成两个大类：

1. **批处理**，顾名思义，它是对批量的数据进行统一处理的方法。比如我们需要知道一个月里，共享单车的总骑行时长和距离，骑行距离最远的单车等信息，就需要对所有单车的数据做汇总的计算，得出想要的结果。批处理的计算过程可能耗时较长，几分钟、几个小时，甚至几天都有可能。
2. **流处理**，它是对数据流做及时的处理计算。相比批处理，它具有低延时的特点，一般可以达到毫秒甚至微秒的级别。这样可以满足海量数据对于大吞吐量的处理能力的需求。比如温度、湿度数据基于高、低阈值的实时判断，就需要采用流处理的方式。

这里要说明一下，批处理延时高的主要原因，是大量数据需要从低速的磁盘中进行存取。它追求的目标和流处理一样，还是高效地处理和尽量简单地计算任务开发工作。

关于批处理和流处理的具体应用，我还会在进阶篇中为你进行讲解。

第五步：数据应用

数据应用处于整个体系的最顶端，是物联网系统的最终目的。我们关注的核心点是通过数据创造**价值**，而实现的基础是各种**算法**。

接下来，我就围绕价值和算法来讲解一下数据应用。

数据如何产生价值呢?

数据产生价值的方法，概况来说，可以分为 4 种：

1. 可视化
2. 挖掘
3. 预测
4. 控制决策

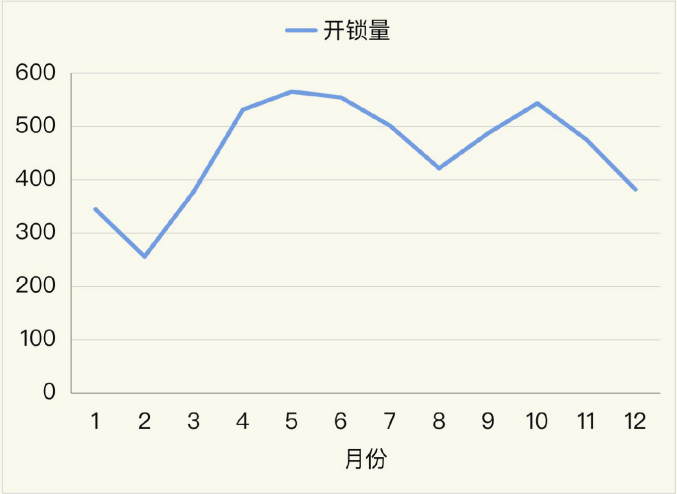
最直接的方法是**可视化**，也可以称为**图表分析**。

可视化可以非常直观地向人们展示数据的含义。比如共享单车的地理位置，如果给用户一个经纬度坐标，那无异于没有给这个信息；而在地图上以光标的方式显示出来，那用户就可以马上知道这个单车相对于自己的方位和距离。

类似的还有以曲线、柱状图表示数据变化趋势，以饼状图表示百分比数据，用网络图表示的事物关系等等。它们都可以为原始数据，和分析结果数据提供理想的展示手段。

比如我们可以查询共享单车近一年的开锁数据，然后基于这个数据，绘制一张开锁量的趋势曲线图。

月份	开锁量
1	345
2	256
3	378
4	532
5	566
6	555
7	502
8	422
9	488
10	544
11	476
12	382



另外，可视化的各种图表也可以作为我们**人工分析数据**的工具。

首先你需要从数据库或者 HDFS （Hadoop 的数据存储格式）中，把数据提取到 Excel 表格（或者 CSV 文件）里面；然后你可以在 Excel 软件中，使用数值计算函数得到总和、平均值等统计数值，或者使用数据透视表等功能得到新的统计表格。

除了 Excel，我们也可以利用其它图表软件来进行分析，比如 Gephi 和 NetMiner。

比图表分析更进一步的方法是**挖掘分析**。

我们可以基于各类算法来分析数据的规律和关联关系等。这种方法基于各种现成的模型，或者自定义的模型，可以从数据中提取出人工分析无法看出来的、隐藏着的规律和联系。

比如那个经常被提到的啤酒和尿布的例子，美国沃尔玛超市通过分析销售数据发现啤酒和尿布经常被一起购买，所以调整货架，将啤酒和尿布放在一起来促进销量（有说法认为，这是由于美国年轻的爸爸出门买尿布的时候，经常顺便买几瓶啤酒犒劳自己）。如果不是挖掘出这个隐藏的联系，我们很难凭直觉得出这个结论，这也体现了挖掘分析的价值。

还有一种进阶的方法是**预测分析**。

随着积累的数据和经验越来越多，我们可以用算法来构建预测模型。当新收集的数据输入给模型时，我们就可以预测出未来的状况。

还是以共享单车做例子，当单车连接到网络平台之后，共享单车出现了。走在路上，掏出兜里的手机，你就可以打开一辆单车的锁，自由骑行；到达目的地后，锁上单车，计费马上完成；你还可以在手机上查看这次骑行的花费时间和骑行速度，甚至计算出你“燃烧的卡路里”。

为什么大部分时候，共享单车可以恰巧出现在你需要它的地方呢？

因为共享单车的系统平台有充足的信息，可以分析出城市的热点区域在哪里，并且预测出第二天每个区域的单车需求量。运营人员正是基于这些数据，提前确定在什么时候、什么地点、投放多少辆车。

最后一种方法是**控制决策**。

物联网设备可能具备**执行器**，物联网系统可以基于业务目标控制这些执行器。而控制决策是采集的数据基于算法计算得到的。

比如在农业中，有经验的种植能手会观察日照、土壤等条件，然后确定施肥、灌溉的节奏。那智慧农业呢，就需要在系统中把这种经验转化为**专家系统**。只要获得光照强度、水分、土壤成分等数据，你就可以根据专家系统的计算结果，控制水泵和卷帘等设备。

数据分析可以使用的算法有哪些?

挖掘分析、预测分析和控制决策要用到统计分析、机器学习，包括人工智能的各种算法。这些算法分为监督学习和非监督学习两类。

1. **监督学习算法**需要我们像尽心尽责的父母，明确告诉它“什么是正确的”，“什么是错误的”。
2. **非监督学习算法**面对的是“放羊式”父母，需要自己将数据中的“异常值”区分出来。

常见的监督学习算法有**决策树**和 **SVM**（支持向量机）算法等。决策树类似我上面描述的专家系统；SVM 算法是一种分类算法，也可以用于线性和非线性回归问题，比如基于骑行习惯构建用户的画像。

非监督学习算法有 **K-means 算法**等。K-means 算法是聚类算法，比如，基于共享单车的使用次数和时间段，我们可以区分出故障车和正常单车。

这里，我没有提到**强化学习**、**半监督学习**等概念，毕竟它们不是这一讲的重点，简单的两句介绍可能让你增加更多疑惑。总之，使用算法的一个原则是，你先要分析数据的**特征**，掌握大体的**倾向**，然后结合你要实现的**目标**，再选择合适的算法。

小结

总结一下，这一讲里，我展开介绍了数据应用涉及的分析方法和相关算法。

1. 数据的分析方法可以分为：可视化、挖掘分析、预测分析和控制决策；
2. 具体的分析中会用到统计分析、机器学习和人工智能的各种算法，这些算法大体可以分为监督学习和非监督学习两类。

3. 算法选择的一个重要原则是，先使用一些可视化的工具分析数据的特征，掌握数据的大体倾向，然后结合业务目标或者控制目标，选择合适的算法。

数据分析和应用是物联网系统中数据流的最后一站。通过数据分析，海量的数据为用户带来了巨大的价值。

数据的价值，首先当然是商业价值，比如超市中的电子价签，取代之前的纸质价签，超市的店面运营减少了很大的人力投入，而且提高了促销活动的效率，降低了价格标错的可能。但这不是全部，物联网数据的有效利用还会带来巨大的社会效益。

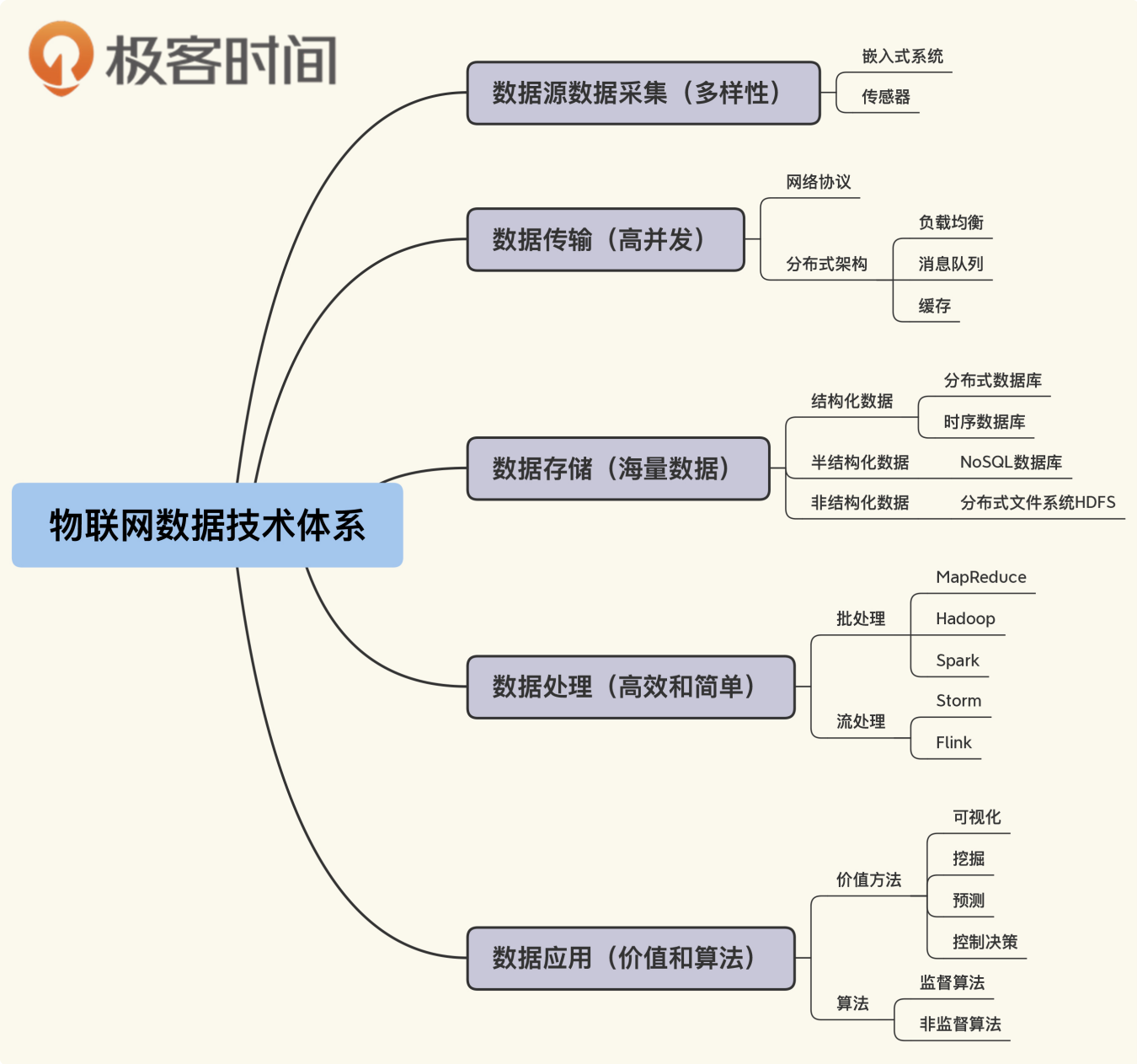
为了讲清楚数据分析应用的方法，我还从数据的角度，对物联网系统再次做了一个梳理，整理出了物联网系统中数据的技术体系。它可以分为 5 个部分，分别是数据源数据采集、数据传输、数据存储、数据处理和数据应用。

可能你也发现了，物联网的数据技术体系和大数据的体系有很多相似的地方。确实是这样，物联网的数据系统本身也是一个大数据系统，必然会用到很多传统的大数据技术，比如数据存储、数据处理相关的技术。

但是物联网系统也有自身的特点，所以在数据采集上会涉及嵌入式系统和传感器的技术；在数据传输中会用到特殊的网络协议；在数据应用中需要关注控制决策相关的技术。这些是你需要注意区分的地方。

总之，在物联网中，海量的数据将会远远超过当前互联网的数据规模。这必然对数据的存储、读取、分析处理技术提出新的挑战。相关的技术呢，我会在进阶篇中继续做深入的剖析。

这里，我再提供一个物联网数据技术的思维导图，供你参考。



思考题

最后，给你留个思考题吧。

在这一讲中，我提到了数据处理可分为批处理和流处理，并简单介绍了它们各自的特点。除了课程中的例子，你认为它们还适合哪些场景呢？

欢迎在留言区写一写你的想法，和我交流一下。也欢迎你将这节课分享给你的朋友一起学习。

提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 03 | 网络协议：设备使用什么语言与互联网对话？

下一篇 05 | 系统实例：怎样设计一个简易物联网系统？

精选留言 (12)

写留言

没有极客精神的极客
2020-11-16

批处理是离线数据分析，流处理是实时数据分析。

展开

7

夜路破晓
2020-11-16

因为目前主要负责数据这块，所以学习本篇的时候主要精力放在本职工作与职业规划之间寻求可能的发展路径方面，明确后期学习的重点方向和内容。给了自己5年时间进入物联网圈子，就像之前讲的，希望在这个圈子能找到自己的一亩三分地扒食糊口顺带做点自己喜欢的事情。在如今“奔四”的路上搭进去一半时间来做这件事，方向、时间、决心都具备了，接下来就剩下脚踏实地干就完了。

展开

作者回复: 赞，加油

5

Geek_41694b
2020-11-17

郭老师，您好，方便的话能不能在每篇文章下面推荐一些文章或是书籍，市面上上的资源质量参差不齐，谢谢

作者回复: 我后面可以考虑在加餐推荐几本书。

1 3

YueShi
2020-11-17

使用例子感觉可以加上《城市大脑》这个东西

据说之前杭州堵成狗了，现在开车大部分情况下是一路绿灯

背后就是一套庞大的物联网系统去支撑

展开

作者回复: 赞，确实是有相关的系统，我在云栖大会看到过这个介绍。

1 3

Sceneryback
2020-11-19

目前携号转网都是整点生效的，估计是整点进行批处理，大概是实时处理能力不足吧😂

作者回复: 哈哈，活学活用了

2

加油加油
2020-11-16

老师您好，请问有没有书籍或者资料更详细的指导物联网平台系统的建设？

作者回复: 这方面的书籍还比较少，平台除了与设备相关的部分，很多还是和互联网一样，你可以分析一下自己真正需要掌握的知识，然后专门去学习，这样效率高。因为即使有这样的书也是分成很多部分，比如设备接入、消息队列、数据存储、数据处理、Web业务系统来讲，还是需要找到你不了解的部分，深入去掌握。

1

marquis
2020-11-25

说到批处理和流处理让我想到两个场景：
1. 在工作中，无论是项目进行时，或者用户体验地图探索，对于一些突发的情况，先着手处理，这个过程是流处理；同时记录下突发事件的类型，等一段时间后复盘，可以看突发

事件更多出现在哪类事情的哪个环节上行，就对那个环节进行探索看是否有改进，这个行为像是批处理。...

展开

莹

2020-11-25

批处理看起来更适合做一些无所谓延时的数据处理，比如根据海量数据制作数据分析图表；流处理更适合一些需要实时分析得出结果的数据处理，比如瞬间图像识别。亚马逊仓库中安装了红外摄像头，用来实时监测员工之间的距离（好像还有测体温），这种就要求实时在显示屏上显示结果，我觉得是流处理。

展开

bulb

2020-11-18

个人对批处理和流处理的一些粗浅理解
批处理，统计汇总性的信息，发现海量数据直接的整体特性和内在管理
流处理，更有针对性，针对单个数据源做实时的判断
不知道理解的对不对

展开

作者回复: 为留言点赞
说回问题，批处理和流处理主要是数据处理方式的不同，流处理一样面临海量的并发数据

独舞

2020-11-17

我们目前项目正用到老师说的这些知识，想快一些消化吸收这些知识，一周3篇的更新太少了，我们一般都不是几天一篇去消化的，可以先通读整体，再重新精读每一个章节，希望老师采纳，非常感谢。

作者回复: 确实办不到，有问题可以沟通群里和我交流哈

Cy23

2020-11-16

要学的东西很多啊，了解整体，需要自我完善分支，视乎对数据处理有了些感觉。

作者回复: 嗯，先建立框架，最后要找到自己主攻的方向。

1

Geek_41694b

2020-11-16

如果是已经全部写好的话，可以一周多更新几篇，感觉一周三篇不怎么过瘾，

作者回复: 还是需留出时间给大家去消化的。学习一定要基于这些内容去扩展，去加深理解，否则很容易停留在只是知道的阶段。你可能对这些内容很了解，那很好。